Estimation of Two-Sided Exponential Distribution Using the Maximum Likelihood Function

I. Akeyede[1]; M. Kaka[2]; and H.R. Bakari[2]

[1]Department of Statistics, Federal University Lafia, PMB 146 Lafia, Nigeria.
[2]Department of Mathematical Sciences, University of Maiduguri, Maiduguri, Nigeria.

*E-mail: akeyede.imam@science.fulafia.edu.ng
kakamodu05@gmail.com
harunbakari@gmail.com

ABSTRACT

This article provides a tutorial exposition of maximum likelihood estimation (MLE). MLE is of fundamental importance in the theory of inference and is a basis of many inferential techniques in Statistics, unlike least squares estimation (LSE), which is primarily a descriptive tool. In this paper, we provide a simple, intuitive explanation of the method so that the reader can have a grasp of some of the basic principles. We hope the reader will apply the method in his or her mathematical modeling in area widely available MLE-based analyses can be performed on data, thereby extracting as much information and insight as possible into the underlying mental process under investigation.

(Keywords: exponential distribution, MLE, LSE)

INTRODUCTION

In mathematical modeling, such hypotheses about the structure and inner working of the behavioral process of interest are stated in terms of parametric families of probability distributions called models. The goal of modeling is to deduce the form of the underlying process by testing the viability of such models. Once a model is specified with its parameters, and data have been collected, one is in a position to evaluate its goodness of fit, that is, how well it fits the observed data. Goodness of fit is assessed by finding parameter values of a model that best fits the data a procedure called parameter estimation.

There are two general methods of parameter estimation. They are least-squares estimation (LSE) and maximum likelihood estimation (MLE). The former has been a popular choice of model fitting in Statistics and is tied to many familiar statistical concepts such as linear regression, sum of squares error, proportion variance accounted for (i.e.: r^2), and root mean squared deviation. LSE, which is unlike MLE requires no or minimal distributional assumptions (non-parametric), is useful for obtaining a descriptive measure for the purpose of summarizing observed data, but it has no basis for testing hypotheses or constructing confidence intervals.

On the other hand, MLE is not as widely recognized among modelers in Statistics, but it is a standard approach to parameter estimation and inference in statistics. MLE has many optimal properties in estimation: sufficiency (complete information about the parameter of interest contained in its MLE estimator); consistency (true parameter value that generated the data recovered asymptotically, i.e., for data of sufficiently large samples); efficiency (lowest-possible variance of parameter estimates achieved asymptotically); and parameterization invariance (same MLE solution obtained independent of the parametrization used).

In contrast, no such things can be said about LSE. As such, most statisticians would not view LSE as a general method for parameter estimation, but rather as an approach that is primarily used with linear regression models. Further, many of the inference methods in statistics are developed based on MLE. For example, MLE is a prerequisite for the chi-square test, the G-square test, Bayesian methods, inference with missing data, modeling of random effects, and many model selection criteria such as the Akaike information criterion (Akaike, 1973) and the Bayesian information criteria (Schwarz, 1978).

MODEL SPECIFICATION

The model specifications are presented as follows:

Probability Density Function

From a statistical standpoint, the data vector $y = (y_1, y_2 \ldots y_m)$ is a random sample from an unknown population. The goal of data analysis is to identify the population that is most likely to have generated the sample. In statistics, each population is identified by a corresponding probability distribution. Associated with each probability distribution is a unique value of the model's parameter. As the parameter changes in value, different probability distributions are generated. Formally, a model is defined as the family of probability distributions indexed by the model's parameters.

Let $f(y/w)$ denote the probability density function (PDF) that specifies the probability of observing data vector y given the parameter w. Throughout this paper we will use a plain letter for a vector (e.g., y) and a letter with a subscript for a vector element (e.g., yi). The parameter $w = (w_1, w_2 \ldots wy_m)$ is a vector defined on a multi-dimensional parameter space. If individual observations, yi's, are statistically independent of one another, then according to the theory of probability, the PDF for the data $y = (y_1, y_2 \ldots y_m)$ given the parameter vector w can be expressed as a multiplication of PDFs for individual observations,

$$f(y = (y_1, y_2 \ldots y_m)/w) = f_1(y_1/w) f_2(y_2/w) \ldots f_n(y_m/w) \quad \textbf{(1)}$$

Likelihood Function

Given a set of parameter values, the corresponding PDF will show that some data are more probable than other data.

Accordingly, we are faced with an inverse problem: Given the observed data and a model of interest, find the one PDF, among all the probability densities that the model prescribes, that is most likely to have produced the data. To solve this inverse problem, we define the likelihood function by reversing the roles of the data vector y and the parameter vector w in $f(y/w)$; that is:

$$L(w/y) = f(y/w). \quad \textbf{(2)}$$

Thus, $L(w/y)$ represents the likelihood of the parameter w given the observed data y; and as such is a function of w: For the one-parameter binomial for example, the likelihood function for $y = 7$ and $n = 10$ is given by:

$$L(w/n = 10, y = 7) = f(y = 7/n = 10, w)$$

$$= \frac{10!}{7!\,3!} w^7 (1 - w)^3, \quad (0 \le w \le 1). \quad \textbf{(3)}$$

There exist an important difference between the PDF $f(y/w)$. and the likelihood function $L(w/y)$. As illustrated above, the two functions are defined on different axes, and therefore are not directly comparable to each other. Specifically, the PDF is a function of the data given a particular set of parameter values, defined on the data scale. On the other hand, the likelihood function is a function of the parameter given a particular set of observed data, defined on the parameter scale.

MAXIMUM LIKELIHOOD ESTIMATION

Once data have been collected and the likelihood function of a model given the data is determined, one is in a position to make statistical inferences about the population, that is, the probability distribution that underlies the data. Given that different parameter values index different probability distributions, we are interested in finding the parameter value that corresponds to the desired probability distribution.

The principle of maximum likelihood estimation (MLE), originally developed by R.A. Fisher in the 1920s, states that the desired probability distribution is the one that makes the observed data ''most likely,'' which means that one must seek the value of the parameter vector that maximizes the likelihood function $L(w/y)$. The resulting parameter vector, which is sought by searching the multi-dimensional parameter space,

is called the MLE estimate, and is denoted by $w_{MLE} = (w_{1,MLE}, w2_{,MLE,\ldots,} w_{K,MLE})$.

To summarize, maximum likelihood estimation is a method to seek the probability distribution that makes the observed data most likely.

Likelihood Equation

MLE estimates need not exist nor be unique. In this section, we show how to compute MLE estimates when they exist and are unique. For computational convenience, the MLE estimate is obtained by maximizing the log-likelihood function; $ln\ L(w/y)$. This is because the two functions, $ln\ L(w/y)$ and $L(w/y)$; are monotonically related to each other so the same MLE estimate is obtained by maximizing either one. Assuming that the log-likelihood function, $ln\ L(w/y)$; is differentiable, if w_{MLE} exists, it must satisfy the following partial differential equation known as the likelihood equation:

$$\frac{\partial ln\ L(w/y)}{\partial w_i} = 0,$$

at $w_i = w_{i,MLE} \forall$ i = 1,2, …, k.

This is because the definition of maximum or minimum of a continuous differentiable function implies that its first derivatives vanish at such points. The likelihood equation represents a necessary condition for the existence of an MLE estimate. An additional condition must also be satisfied to ensure that $ln\ L(w/y)$ is a maximum and not a minimum, since the first derivative cannot reveal this.

To be a maximum, the shape of the log-likelihood function should be convex (it must represent a peak, not a valley) in the neighborhood of w_{MLE}: This can be checked by calculating the second derivatives of the log-likelihoods and showing whether they are all negative at:

$$w_i = w_{i,MLE} \quad \forall\ \mathrm{i} = 1,2,\ldots,\mathrm{k}, \quad \frac{\partial^2 ln\ L(w/y)}{\partial \mathrm{w}_i^2} < 0.$$

Maximum Likelihood Estimation of a Two-Sided Exponential Probability Distribution Function

In this section, we present an application of maximum likelihood estimation by using an illustrative example, in doing this the exponential model and the power model will be use and the models are as define as:

Power model:

$$f(w,t) = w_1 t^{-w_2}, \quad (w_1, w_2 > 0)$$

$$= \frac{n!}{(n-x)!x!}(w_1 t_i^{-w_2})^{x_i}(1 - w_1 t_i^{-w_2})^{(n-x_i)} \qquad \textbf{(4)}$$

Exponential model:

$$f(w,t) = w_1 \exp(-w_2 t), \quad (w_1, w_2 > 0)$$

$$= \frac{n!}{(n-x)!x!}(w_1 \exp(-w_2 t))^{x_i}(1 - w_1 \exp(-w_2 t))^{(n-x_i)} \qquad \textbf{(5)}$$

$$where\ x_i = 0,1,\ldots,n, \quad i = 1,2,\ldots,m.$$

Now, assuming that $x_i's$ are statistically independent of one another, the desired log-likelihood function for the power model is given by:

$$lnL(w = (w_1, w_2)|n,x) = ln[f(x_1|n,w).f(x_2|n,w)\ldots f(x_m|n,w)]$$

$$= \sum_{i=1}^{m} lnf(x_i|n,w)$$

$$= \sum(x_i \ln(w_1 t_i^{-w_2}) + (n - x_i)\ln(1 - w_1 t_i^{-w_2}) + lnn! - \ln(n - x_i)! - lnx_i!) \qquad \textbf{(6)}$$

This quantity is to be maximized with respect to the two parameters, w_1 and w_2. It is worth noting that the last three terms of the final expression in the above equation,

(i.e., $\ln n! - \ln(n - x_i)! - \ln x_i!$),

do not depend upon the parameter vector, thereby do not affecting the MLE results. Accordingly, these terms can be ignored, and their values are often omitted in the calculation of the log-likelihood. Similarly, for the exponential model, its log-likelihood function can be obtained from above by substituting $w_1 \exp(-w_2 t_i)$ for $w_1 t^{-w_2}{}_i$.

Information Matrix

We can obtain a 2×2 square information matrix for the two-sided exponential distribution presented.

$$I = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$$

For our model, all the second derivatives exist. Hence we can obtain:

$$I_{11} = 0$$

$$I_{12} = I_{21} = -te^{-w_2 t}$$

$$I_{22} = w_1 t^2 e^{-w_2 t}$$

Hence,

$$I = \begin{bmatrix} 0 & -te^{-w_2 t} \\ -te^{-w_2 t} & w_1 t^2 e^{-w_2 t} \end{bmatrix} \quad \textbf{(7)}$$

In illustrating MLE, we used a data set from Murdock (1961). In this experiment subjects were presented with a set of words or letters and were asked to recall the items after six different retention intervals, ($t_1,\ldots,t_6$) = (1; 3; 6; 9; 12; 18) in seconds and thus, m = 6: The proportion recall at each retention interval was calculated based on 100 independent trials (i.e. n = 100) to yield the observed data ($y_1,\ldots,y_6$) = (0.94; 0.77; 0.40; 0.26; 0.24; 0.16); from which the number of correct responses, x_i, is obtained as $100y_i$, i = 1,…, 6: In Figure 1, the proportion recall data are shown as squares.

The curves in Figure 1 are best fits obtained under MLE. Table 1 summarizes the MLE results, including fit measures and parameter estimates, and also includes the LSE results, for comparison.

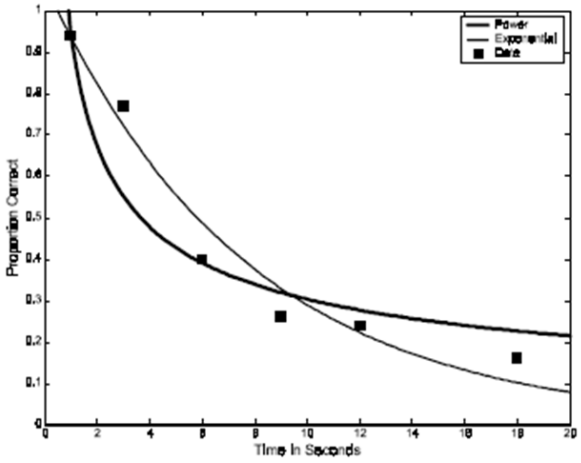

Figure 1: Modeling Forgetting Data. Squares represent the data in Murdock (1961). The Thick (respectively, thin) are best fits by the power (respectively, exponential) models.

Table 1: Summary Fits of Murdock (1961) Data for the Power and Exponential Models under the Maximum Likelihood Estimation (MLE) Method and the Least Squares Estimation (LSE) Method.

	MLE		LSE	
	Power	Exponential	Power	Exponential
Loglik/SSE (r^2)	−313.37 (0.886)	−305.31 (0.963)	0.0540 (0.894)	0.0169 (0.967)
Parameter w_1	0.953	1.070	1.003	1.092
Parameter w_2	0.498	0.131	0.511	0.141

Note: For each model fitted, the first row shows the maximized log-likelihood value for MLE and the minimized sum of squares error value for LSE. Each number in the parenthesis is the proportion of variance accounted for (i.e. r^2) in that case. The second and third rows show MLE and LSE parameter estimates for each of w_1 and w_2. The above results were obtained using Matlab code described in the appendix.

The results in Table 1 indicate that under either method of estimation, the exponential model fit better than the power model. That is, for the former, the loglikelihood was larger and the SSE smaller than for the latter. The same conclusion can be drawn even in terms of r^2. Also note the appreciable discrepancies in parameter estimate between MLE and LSE.

These differences are not unexpected and are due to the fact that the proportion data are binomially distributed, not normally distributed. Further, the constant variance assumption required for the equivalence between MLE and LSE does not hold for binomial data for which the variance $\sigma^2 = np(1-p)$, depends upon proportion correct p.

MLE INTERPRETATION

What does it mean when one model fits the data better than does a competitor model? It is important not to jump to the conclusion that the former model does a better job of capturing the underlying process and therefore represents a closer approximation to the true model that generated the data. A good fit is a necessary, but not a sufficient, condition for such a conclusion. A superior fit (i.e., higher value of the maximized loglikelihood) merely puts the model in a list of candidate models for further consideration. This is because a model can achieve a superior fit to its competitors for reasons that have nothing to do with the model's fidelity to the underlying process. For example, it is well established in statistics that a complex model with many parameters fits data better than a simple model with few parameters, even if it is the latter that generated the data. The central question is then how one should decide among a set of competing models. A short answer is that a model should be selected based on its generalizability, which is defined as a model's ability to fit current data but also to predict future data.

CONCLUDING REMARKS

This article provides a tutorial exposition of maximum likelihood estimation. MLE is of fundamental importance in the theory of inference and is a basis of many inferential techniques in statistics, unlike LSE, which is primarily a descriptive tool. The intended audiences of this tutorial are researchers who practice mathematical modeling of cognition but are unfamiliar with the estimation method. Unlike least-squares estimation which is primarily a descriptive tool, MLE is a preferred method of parameter estimation in statistics and is an indispensable tool for many statistical modeling techniques, in particular in non-linear modeling with non-normal data.

In this paper, we provide a simple, intuitive explanation of the method so that the reader can have a grasp of some of the basic principles. We hope the reader will apply the method in his or her mathematical modeling efforts so a plethora of widely available MLE-based analyses can be performed on data, thereby extracting as much information and insight as possible into the underlying mental process under investigation.

REFERENCES

1. Annette, J. D. 2001. *Introduction to Generalized Linear Models. 2nd edition*. CRC Press: Boca Raton, FL.

2. Bain, L. J. 1978. *Statistical Analysis of Reliability and Life-Testing Models*. Decker: New York, NY.

3. Bain, L.J. and M. Engelhart. 1989. *Introduction to Probability and Mathematical Statistics*. PWS: Kent, MA.

4. Balkema, A., and L. de Haan. 1974. "Residual Llifetime at Great Age". *Annals of Probability*. 2: 792-804.

5. Barlow, R.E., and F. Proschan. 1975. *Statistical Theory of Reliability and Life Testing*. Holt, Rinehart, & Winston: New York, NY.

6. Dodson, B. 1994. *Weibull Analysis*. ASQC: Milwaukee, WI.

7. Embrechts, P., C. Klüppelberg, and T. Mikosch. 1997. *Modeling Extremely Events for Insurance and Finance*. Spring Verlag: Berlin, Germany.

8. Harville, D.A. 1977. "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems". *Journal of the American Statistical Association*. 72: 320-340.

9. Hogg, R.V. and A.T. Craig. 1970. *Introduction to Mathematical Statistics*. Macmillan: New York, NY.

10. Kasumu, R.B. 2002. *Elements of Statistical Inference*. Jas Publishers.

11. Lee, E.T. 1980. *Statistical Methods for Survival Data Analysis: Lifetime Learning*. Belmont, CA.

12. Lieblein, J. 1955. “On Moments of Order Statistics from the Weibull Distribution”. *Annals of Mathematical Statistics*. 26: 330-333.

13. Meeker, W.Q. and L.A. Escobar. 1998. *Statistical Methods for Reliability Data*. John Wiley & Sons: New York, NY.

14. Murdock, B. B., Jr. 1961. “The Retention of Individual Items”. *Journal of Experimental Psychology*. 62: 618-625.

15. Pickands, J. 1975. “Statistical Inference using Extreme Order Statistics”. *Annals of Statistics*. 3: 119-131.

16. Wilks, S.S. 1946. Mathematical Statistics. Princeton University Press: Princeton, NJ.

SUGGESTED CITATION

Akeyede, I., M. Kaka, and H.R. Bakari. 2021. “Estimation of Two-Sided Exponential Distribution Using the Maximum Likelihood Function”. *Pacific Journal of Science and Technology*. 22(1): 24-29.